MESRGAN on Sign Language image Recognition

Chan Jyh Huah (A0178199H), Huang Fuxing (A0163461J)
Tan Chee Wei (A0179723U), Woo Chee Yoong (A0178344X)

Institute of Systems Science, 25 Heng Mui Keng Terrace, Singapore 119615

Submitted Aug 2019

Abstract

American Sign Language (ASL) is a complete sign language system that is widely used by deaf individuals in the United States and the English-speaking part of Canada. An automatic ASL recognition system is highly desirable for solving the constraint on communicating with deaf people for non-sign-language speakers. In this paper, we proposed a novel architecture, MESRGAN, which combines ESRGAN image optimization and a high performing neural network activation function, GELU to achieve a higher accuracy in sign language recognition using the ASL dataset.

Keywords: ASL, sign language recognition, GELU, ESRGAN, CNN

1. Introduction

American Sign Language (ASL) is a complete sign language system that is widely used by deaf individuals in the United States and the English-speaking part of Canada. An automatic ASL recognition system is highly desirable for solving the constraint on communicating with deaf people for non-sign-language speakers. One approach would be to interpret gesture signs into humanoid or machine decipherable text [1].

1.1 Dataset

The dataset format is patterned from the classic MNIST dataset, which is a popular benchmark for image-based machine learning. The training data (27,455 cases) and test data (7,172 cases) will be applied to K-Fold validation using train-test split of 70:30. Each image is represented by 28x28 pixel and in grayscale with values ranging between 0-255. There will be a one-to-one mapping for each alphabetic letter A-Z, with the labels set to 0-25, excluding 9-J & 25-Z because the sign for these two alphabets require movement.

Figure 1 ASL Alphabets

1.2 Related works

A CNN model with five layers and hyperparameters optimization approach was proposed in [2] to be applied to the sign language recognition problem domain with an accuracy close to 90%. Another paper [3] focuses on experimenting with different segmentation approaches and unsupervised

learning algorithms to create an accurate sign language recognition model using RGB images collected from a Microsoft Kinect and fed into the autoencoder to extract features. The method achieved a classification accuracy of 98% trained on a subset of 10 alphabets instead of the full 24. A third reference paper [4] extract features by using a depth contrast feature based per-pixel classification algorithm with a hierarchical mode-seeking method to localize hand joint positions under kinematic constraints and trained with a Random Forest (RF) classifier to recognize ASL signs using the joint angles which was able to achieve above 90% accuracy in recognizing 24 static ASL alphabet signs.

2. Our Approach

2.1 ESRGAN Image Optimization

As there are many models and approach to solve this problem, we choose to focus on developing a novel approach which could provide substantial improvement in the accuracy of the model by combining image dataset optimization and modelling. To work on the first part of our approach, we refer to the method proposed in [5] which further enhance the visual quality by adopting a deeper model using Residual-in-Residual Dense Block (RRDB) without batch normalization layers, employing Relativistic average GAN instead of the vanilla GAN, which learns to judge whether one image is more realistic than another, guiding the generator to recover more detailed textures and enhanced the perceptual loss by using the features before activation, which offer stronger supervision and thus restore more accurate brightness and realistic textures.

2.2 GELU neural network activation function

The second part of our approach looks at improving the model architecture to further improve on the accuracy. For this we refer to the approach proposed in [6] to optimize on the neural network activation function which randomly applies zero map to a neuron's input and weigh inputs by their magnitude rather than by their sign as in RELU's. This activation function shows overall performance improvement across computer vision, natural language processing and speech tasks. We often want a deterministic decision from a neural network, and this gives rise to our new nonlinearity. The nonlinearity is the expected transformation of the stochastic regularizer on an input x, which is $\Phi(x) \times Ix + (1 - \Phi(x)) \times 0x = x\Phi(x)$. Loosely, this expression states that we scale x by how much greater it is than other inputs. Since the cumulative distribution function of a Gaussian is often computed with the error function, we define the Gaussian Error Linear Unit (GELU) as

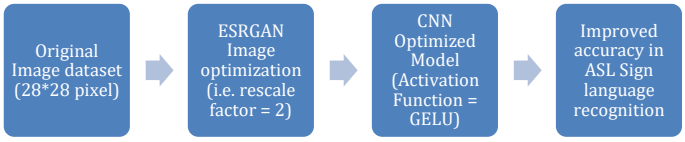
$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x).$$

We can approximate the GELU with

$$0.5x(1 + \tanh[\sqrt{2/\pi}(x + 0.044715x^3)])$$

2.3 MESRGAN

With the above two approaches, we can now build the workflow to construct our model architecture which we named as MESRGAN (More Enhanced SRGAN) as follows:



Given the original image dataset, we can leverage on ESRGAN to create a higher res image dataset and this can later be trained on a CNN model using the GELU activation function to achieve higher accuracy.

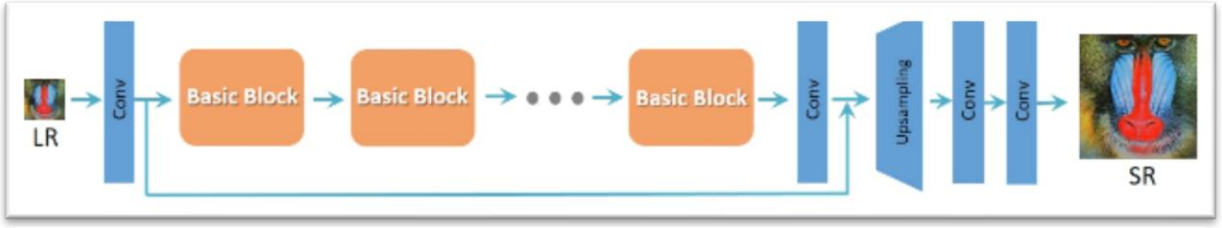


Figure 2 Enhanced SRGAN for image dataset preprocessing

3. Experiments

A seven K-fold cross validation method is used to build the model using the training dataset which are later used to score the model accuracy on the testing

dataset. The CNN baseline model which we use in our architecture is referenced from [7] which is a popular machine learning website hosting the dataset. The ESRGAN image optimization which we used in this experiment is set to the default setting with a rescale factor of 2x and with the GELU activation function. To cover the different variations, we ran the experiment a total of three times with the baseline model, the ESRGAN image optimization only and our proposed model, MESRGAN.

3.1 Result

For benchmark, we have included the accuracy of the referenced three papers to set a rough ballpark estimates of where the accuracy should land. The accuracy shown on the second paper is at 98% due to the reduced scope of the training class to ten instead of the full 24 available alphabets. The ESRGAN only model shows an improvement of near 7% against the baseline CNN model and our proposed method with the GELU activation function can be seen to further optimized the accuracy for an additional 1% gain.

	APPROACH	ACCURACY
1	Convolutional Neural Network	89.32%
2	Unsupervised Feature Learning + CNN	98.0%
3	Random Forest	90%
4	**Baseline CNN**	85.72%
5	**ESRGAN + CNN**	92.47%
6	**Our Method (MESRGAN)**	93.63%

3.2 Future Work

Although our method shows an improvement over the baseline model, the accuracy shown are still far from the state-of-the-art accuracy achieved using other modelling methods. Hence additional work can be done to further optimized the current configuration to max out the ESRGAN image optimization such as testing with different scaling factor or replacing the CNN model with other models derived from transfer learning such as VGGNET16, ResNet50 or DCNN. Finally to improve the value of the work to close the gap to practical application in the real world, we would like to implement the model on video dataset instead of static image which can allow for real time sign language recognition.

4. Conclusion

In this paper, we proposed a novel architecture, MESRGAN, which combines ESRGAN image optimization and a high performing neural network activation function, GELU to achieve a higher accuracy in sign language recognition using the ASL dataset.

References

1. B M Chethana Kumara, H S Nagendraswamy and R Lekha Chinmayi, “Spatial Relationship Based Features for Indian Sign Language Recognition”, International Journal of Computing, Communications & Instrumentation Engineering, Vol. 3, Issue 2, ISSN 23491469, 2016
2. Zhao, Y., & Wang, L. (2018, November). The Application of Convolution Neural Networks in Sign Language Recognition. In 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP) (pp. 269-272). IEEE.
3. Chen, Justin. Sign Language Gesture Recognition with Unsupervised Feature Learning 2011, cs229.stanford.edu/proj2011/ChenSenguptaSundaram-SignLanguageGestureRecognitionWithUnsupervisedFeatureLearning.pdf.
4. Dong, C., Leu, M. C., & Yin, Z. (2015). American sign language alphabet recognition using microsoft kinect. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 44-52).
5. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 0-0).
6. Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.
7. Bakhova, M. (2019). CNN with Image Augmentation | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/mathemilda/cnn-with-image-augmentation [Accessed 1 Sep. 2019].